Identificación de Entidades con Nombre basada en Modelos de Markov y Árboles de Decisión *

José A. Troyano, Víctor J. Díaz, Fernando Enríquez
Javier Barroso y **Vicente Carrillo**
Universidad de Sevilla
Avda. Reina Mercedes s/n
41012 Sevilla
{troyano,vjdiaz}@lsi.us.es

Resumen: Este artículo presenta un sistema para el reconocimiento de entidades con nombre apoyándonos en dos técnicas clásicas de aprendizaje automático: los modelos de Markov y los árboles de decisión. Se han desarrollado varios sistemas en los que hemos investigado el efecto producido por la inclusión de características que no dependen en exceso del idioma utilizado. Los experimentos se han realizado con el corpus del español distribuido para la tarea de reconocimiento de entidades con nombre del CoNLL 2002.
Palabras clave: Reconocimiento de entidades, Modelos de Markov, Árboles de decisión

Abstract: In this paper we investigate Named Entity Recognition (NER) systems using two well-known classifiers in the machine learning literature: Markov Models and Decision Trees. We have designed several systems to check the impact of introducing different characteristics which have a weak dependence of the language used. We also report the results obtained by our systems on the Spanish corpus provided in the NER Task of ConNLL 2002 conference.
Keywords: Name-Entity Recognition, Markov Models, Decision Trees

1. *Introducción*

Un sistema encargado del reconocimiento de entidades con nombre (NER, *Named Entity Recognition*) persigue delimitar en un texto arbitrario aquellas frases simples que responden de forma directa a preguntas del tipo ¿quién?, ¿dónde?, ¿cuándo? o ¿cuánto?. Por ejemplo, dado el siguiente texto:

> El presidente del Consejo Municipal Social, Luis M. Japón, advirtió en una conferencia en Madrid celebrado el día del Trabajo, que el gobierno pretende a partir de enero subir el precio del tabaco en toda España más de un 20 por ciento.

un sistema NER reconocería un nombre de persona (Luis M. Japón), dos localizaciones (Madrid, España), una organización (Consejo Municipal Social), dos fechas (día del Trabajo, enero) y una cantidad (20 por ciento).

El reconocimiento de estas entidades se considera un paso previo hacia la comprensión automática de un texto, ya que aportan mucha información sobre su contenido. Por tanto, es frecuente que los sistemas NER estén integrados dentro de otros sistemas más complejos que abordan la recuperación de información, la extracción de información o los sistemas de búsquedas de respuestas, por citar tan sólo algunos de ellos.

Las conferencias MUC-6 y MUC-7 (Chinchor, 1998) constituyeron un hito a la hora de especificar lo que se entiende por entidad con nombre y la forma de evaluación de los sistemas NER. Los resultados tan prometedores que se obtuvieron entonces han motivado el desarrollo de sistemas NER cada vez más ambiciosos en los que se valoran otros factores como la capacidad de adaptación a distintos géneros y formatos de textos, el multilingüismo, la limitación en el uso de información externa como diccionarios o *gazetters*, etc.

1.1. Planteamiento del problema

Los sistemas NER pueden ser interpretados como un problema de clasificación en el que dado un texto representado como una secuencia de palabras (o tokens) $\bar{w} = w_1 \dots w_T$

* Parcialmente financiado por el Ministerio de Ciencia y Tecnología (FIT-150500-2002-416).

se desea asociar a cada palabra w_i una etiqueta t_i que determina el tipo de entidad que es. En nuestro caso adoptaremos la ontología propuesta en (Tjong, 2002) donde se consideran las siguientes clase de entidades: PER para personas, LOC para localizaciones, ORG para organizaciones y MISC para entidades misceláneas que no se ciñen a los tres anteriores tipos.

Para que cada palabra disponga de una etiqueta propia utilizaremos la notación BIO (figura 1). La etiqueta de la primera palabra perteneciente a una entidad de tipo XXX será etiquetada mediante B-XXX. Las siguientes palabras, si existen, pertenecientes a la misma entidad serán etiquetadas I-XXX. Finalmente, aquellas palabras del texto que no son consideradas entidades presentarán la etiqueta O. De esta forma, si deseamos reconocer n tipos de entidades, dispondremos de $2n + 1$ etiquetas.

Palabra	Etiqueta
La	O
Delegación	B-ORG
de	I-ORG
la	I-ORG
Agencia	I-ORG
EFE	I-ORG
en	O
Extremadura	B-LOC
transmitirá	O
...	...

Figura 1: Ejemplo de texto etiquetado

Un aspecto importante en todo desarrollo de un sistema NER es poder calibrar su calidad. Para conseguirlo se suelen adoptar medidas que determinen, dado un texto de referencia previamente etiquetado, la divergencia que existe entre las etiquetas propuestas y las del texto. Las medidas más habituales que se aplican son la cobertura C y precisión P definidas mediante:

$$P = \frac{\#\text{ entidades correctas en el análisis propuesto}}{\#\text{ entidades en el análisis propuesto}}$$

$$C = \frac{\#\text{ entidades correctas en el análisis propuesto}}{\#\text{ entidades en el texto}}$$

Es frecuente aportar además una medida de compensación entre la cobertura y precisión denominada medida F_β definida mediante:

$$F_\beta = \frac{(\beta^2 + 1)PC}{\beta^2 PC}$$

2. *Modelos de aprendizaje*

Una vez planteado el reconocimiento de entidades como un problema de clasificación es posible la adopción de métodos de aprendizaje automático supervisado. Estos métodos parten de un texto (*corpus*) de referencia previamente etiquetado (conjunto de entrenamiento) y su objetivo es capturar el conocimiento implícito es dicho conjunto con la idea de aplicarlo sobre textos desconocidos. Una de las metas que nos hemos marcado en este trabajo es explorar qué grado de éxito puede ser alcanzado por un sistema NER utilizando como base dos de los métodos más ampliamente divulgados en la literatura: los modelos de Markov y los árboles de decisión.

2.1. Modelos de Markov

Considerando que el etiquetado de entidades sigue un proceso markoviano de primer orden podemos determinar que dada un texto de T palabras $w = (w_1, \ldots, w_T)$ su etiquetado $\bar{t} = (t_1, \ldots, t_T)$ vendría dado por:

$$\bar{t} = \underset{t_1,\ldots,t_T}{\operatorname{argmax}} \prod_{i=1}^{i=T} P(t_i|t_{i-1})P(w_i|t_i)$$

considerando una etiqueta extra t_0 =O para marcar el principio de etiquetado. El factor $P(t_i|t_{i-1})$ establece la probabilidad de *transición* de una etiqueta a otra. El factor $P(w_i|t_i)$ establece la probabilidades *léxicas* de que una palabra sea emitida por una etiqueta. Las probabilidades léxicas y de transición deben ser estimadas y almacenadas en sendas matrices. Una vez conocidas ambas, se puede aplicar el algoritmo de Viterbi para calcular el vector $\bar{t}$.

Visto lo anterior, el problema fundamental con el que nos enfrentamos es la determinación de ambas matrices. A partir de un corpus de entrenamiento podemos obtener el valor de los parámetros aplicando el estimador de máxima probabilidad. La única dificultad es que la carencia de suficientes ejemplos (dispersión de datos) puede introducirnos demasiados parámetros con valores nulos o muy cercanos a cero. Este problema, más agudo cuanto mayor es el conjunto de etiquetas o palabras, puede ser amortiguado aplicando técnicas de suavizado que reparten

la masa probabilística entre todos las observaciones.

2.2. Árboles de decisión

Los árboles de decisión son clasificadores inspirados en la estrategia *divide y vencerás*. Un árbol de decisión es un tipo especial de árbol donde cada nodo interior se asocia con una pregunta sobre un determinado conjunto de atributos X. Las respuestas a dichas preguntas determinan cada uno de los caminos que parten de cada nodo. Dado un objeto o descrito mediante una tupla de atributos $(x_1, \ldots x_m) \in X$, el proceso de clasificación comienza desde la raíz y desciende por las ramas del árbol según sean las respuestas obtenidas en cada uno de los nodos hasta alcanzar la hoja que determinará su clase.

El problema de la construcción eficiente de árboles de decisión a partir de corpus ha sido objeto de amplio estudio, profundizando en aquellos criterios de división de nodos que favorecen la ganancia de información. Ahora bien, cuando el problema de clasificación es complejo, el árbol de decisión puede tener un tamaño tal que no sea fácil su comprensión. También surgen dificultades cuando abordamos objetos del que no conocemos completamente todas sus características. A pesar de esto, los árboles de decisión son uno de los clasificadores más ampliamente utilizados en el campo del aprendizaje automático.

2.3. Combinación de clasificadores

En muchas ocasiones es deseable disponer de varias opiniones antes de tomar una decisión. Esta idea es la que respalda a dos técnicas de aprendizaje que se basan en la combinación de distintos clasificadores. Dos de las técnicas más utilizadas son el *boosting* y el *bagging* (Witten y Frank, 2002). Ambas se basan en la generación de distintos clasificadores a partir del mismo conjunto de entrenamiento, pero se diferencian en la manera en la que estos clasificadores son construidos. Las ideas básicas de estas técnicas son:

- *Bagging*: Consiste en la generación de distintos conjuntos de entrenamiento a partir del conjunto original. Para generar cada nuevo conjunto, se muestrea el original eliminando y replicando ejemplos de forma aleatoria. Cada conjunto de entrenamiento dará lugar a un clasificador distinto.
- *Boosting*: En lugar de generar cada modelo de forma independiente como lo hace el *bagging*, el *boosting* es una técnica iterativa que va aprovechando en cada iteración los resultados de la iteración anterior. Cada nuevo modelo se fuerza para que dé mejor respuesta en los ejemplos en los que los modelos anteriores fracasaron. Ello se consigue dando distintos pesos a los ejemplos del conjunto de aprendizaje.

Una vez generados los distintos clasificadores, el resultado final se obtiene por votación (ponderada en el caso del *boosting*).

3. Características utilizadas

Los modelos de Markov tienen muchas ventajas, pero entre ellas no se encuentra una gran flexibilidad para integrar distintas características. En un problema como el del reconocimiento de entidades con nombre, esto supone una limitación ya que alrededor de una entidad se pueden encontrar muchas pistas para determinar sus límites y para clasificarla.

Básicamente se pueden distinguir dos posibles soluciones a la integración de características en los modelos de Markov.

1. Aquellas que incorporan información adicional transformando el corpus de entrenamiento mediante la sustitución de las etiquetas originales por otras que aporten más información. Por ejemplo, en (Rössler, 2002) se utiliza esta técnica para especializar las etiquetas de las palabras que aparecen con más frecuencia antes y después de cada tipo de entidad.
2. Aquellas, como sucede en (Zhou y Su, 2002), que intentan incluir esa información en las palabras, en vez de en las etiquetas. Bajo este enfoque, las probabilidades léxicas pueden ser estimadas teniendo en cuenta características más complejas. Una forma parecida de proceder, esta vez aplicada a la tarea de etiquetado morfosintáctico, la encontramos en (Brants, 2000) donde la probabilidad de las palabras desconocidas se estima en base a la frecuencia de secuencias de letras como, por ejemplo, los sufijos.

Ambas soluciones requieren un mayor volumen de datos de entrenamiento a medida que se incorporan nuevas características.

Las primeras porque amplían el conjunto de etiquetas y aumenta por tanto la dimensión de la matriz de transiciones, las segundas porque amplían la matriz léxica en función del número de posibles combinaciones de características.

En muchos casos la identificación de entidades con nombre plantea un problema de clasificación bastante difícil, por ejemplo la palabra `Sevilla` puede hacer referencia a un lugar, a una organización (el ayuntamiento) o incluso a una persona (un apellido). Ante este tipo de ambigüedades hay muchos elementos que pueden ser relevantes para clasificar adecuadamente, de manera que cuantas más características tengamos en cuenta mejor. Si queremos explorar este camino y añadir un número considerable de características, no podremos hacerlo apoyándonos exclusivamente en los modelos de Markov, por los problemas que plantearía la estimación de los parámetros.

Este planteamiento nos lleva a descargar de trabajo al modelo de Markov, exigiéndole exclusivamente la tarea de identificar las secuencias de palabras que forman parte de una entidad, sin determinar la categoría a la que pertenece. Con ello se resuelven dos problemas de distinta naturaleza, la delimitación y la clasificación, lo que permite aplicar a cada uno de ellos la técnica más apropiada, y utilizar en cada caso también las características más apropiadas.

En las siguientes subsecciones describiremos las diferentes características que hemos probado en nuestros experimentos.

3.1. En la delimitación

En la delimitación de entidades, la inclusión de características, más que añadir información, lo que nos va a permitir es quedarnos exclusivamente con la información que nos hace falta. La técnica que hemos elegido para introducir este tipo de información es la de transformación del corpus (Rössler, 2002). Básicamente esta técnica consiste en modificar tanto el corpus de entrenamiento como el de test de manera que se obtenga un modelo de Markov distinto al que se obtendría con el corpus original. Con ello se pueden conseguir varias cosas:

- modelar un problema de etiquetado distinto al original
- dar más peso a ciertas palabras del vocabulario
- agrupar varias palabras del vocabulario en una única entrada

La figura 1, que presenta un fragmento del corpus original, nos servirá de apoyo a la hora de mostrar las distintas transformaciones que hemos aplicado al corpus. Para que el modelo de Markov se encargue exclusivamente de delimitar las entidades existentes procedemos a transformar el corpus de entrenamiento eliminando de las etiquetas la información relativa a la categoría de las entidades (figura 2).

Palabra	Etiqueta
La	O
Delegación	B-XXX
de	I-XXX
la	I-XXX
Agencia	I-XXX
EFE	I-XXX
en	O
Extremadura	B-XXX
transmitirá	O
...	...

Figura 2: Transformación de etiquetas

La estrategia de simplificar la tarea del modelo de Markov da resultados bastante buenos al detectar entidades que han aparecido en el corpus de entrenamiento pero se comporta bastante mal ante entidades desconocidas. El problema de las palabras desconocidas es bastante común en todo proceso de etiquetado, pero es más crítico aún, si cabe, en la identificación de entidades ya que dichas palabras suelen ser buenas candidatas para formar parte de una entidad. La falta de información que conlleva el desconocimiento de una palabra puede ser compensada con la información que proporcionan las mayúsculas, ya que en español gran parte de los nombres de entidades contienen palabras en mayúsculas y, a la inversa, la mayor parte de las palabras en mayúsculas forman parte de entidades con nombre. Además de las palabras que contienen mayúsculas, hay otro tipo de palabras que pueden ser de ayuda en la detección de nombres de entidades, se trata de aquellas palabras en minúscula que aparecen frecuentemente alrededor o dentro de una entidad.

Tanto la información de las mayúsculas como la de las palabras que aparecen de forma frecuente rodeando o dentro de una entidad ha sido introducida a través de transformaciones del corpus de entrenamiento. A diferencia de la transformación anterior, en este caso no se modifican las etiquetas sino que son las palabras del vocabulario las que se transforman. Se aplican las siguientes reglas:

- Se sustituye cada palabra del corpus por un *token* representativo. Los tokens considerados son: _may_ para palabras que comienzan con mayúsculas, _min_ para minúsculas, _tmay_ para palabras sólo con mayúsculas, _mypto_ para una mayúscula seguida de un punto, _abrev_ para abreviaturas en general, y _1apal_ para aquellas palabras que comienzan una frase.
- Se excluyen de esta transformación aquellas palabras que aparecen frecuentemente alrededor o dentro de una entidad.

Palabra	Etiqueta
La	O
may	B-XXX
de	I-XXX
la	I-XXX
may	I-XXX
tmay	I-XXX
en	O
may	B-XXX
min	O
...	...

Figura 3: Transformación de etiquetas y palabras

Dado que esta última transformación afecta al vocabulario debe aplicarse tanto al corpus de entrenamiento como al de test. La figura 3 muestra el aspecto del fragmento del corpus original (figura 1) tras la aplicación de las transformaciones de etiquetas y vocabulario.

3.2. En la clasificación

Tras delimitar las entidades disponemos de un texto en el que han sido señaladas las posibles entidades sin especificar a qué clase pertenecen. Para cada una de estas entidades se genera un vector de características que son utilizadas para clasificar.

Los ejemplos de entrenamiento son generados aplicando el mismo proceso de extracción de vectores de características a las entidades presentes en el corpus de entrenamiento.

Las características utilizadas en los experimentos se agrupan en las siguientes categorías:

1. Ortográficas: A pesar de que parte de esta información ha sido ya utilizada en el proceso de delimitación puede aún ser útil a la hora de discriminar la categoría de una entidad. Se tiene en cuenta si una entidad contiene palabras que comienzan en mayúsculas, totalmente en mayúsculas, dígitos y números romanos. Se incluyen también en este grupo otras características como la longitud en palabras de una entidad, la posición relativa dentro de la frase y si contiene comillas o no.
2. Sufijos: Se calculan a partir del corpus de entrenamiento los sufijos de dos y tres letras relevantes para cada categoría.
3. Contextos: En una ventana de tres palabras alrededor de una entidad se calculan las palabras relevantes para cada categoría según los ejemplos del corpus de entrenamiento.
4. Palabras significativas: Para cada categoría se calcula el conjunto de palabras significativas eliminando palabras huecas y en minúsculas de los ejemplos del corpus de entrenamiento.
5. Listas externas: Se comprueba si una entidad presenta alguna palabra perteneciente a una serie de listas generadas con información externas al corpus de entrenamiento. Se tienen en cuenta listas de nombres, apellidos, países, ciudades y disparadores internos de cada categoría (por ejemplo *calle* para la categoría LOC).

En las características 2, 3 y 4 se dispone de una serie de listas generadas automáticamente a partir del corpus de entrenamiento. Todas los elementos de una lista no son igualmente relevantes ya que algunos serán muy frecuentes en el corpus de entrenamiento y otros lo serán menos. En nuestros ex-

perimentos aprovechamos esa información a la hora de computar la característica de una entidad, dando más peso a los elementos más frecuentes.

4. Experimentos

Todos los experimentos se han realizado con el corpus del español distribuido para la tarea de reconocimiento de entidades con nombre del CoNLL 2002 (Tjong, 2002). Dicha distribución consta de un corpus de entrenamiento de 264715 tokens y 18794 entidades, un test A (52923 tokens y 4315 entidades) que ha sido utilizado durante el desarrollo y un test B (51533 tokens y 3558 entidades) utilizado exclusivamente en las pruebas finales.

4.1. Experimento 1

En la tabla 1 se muestran los resultados de nuestro primer experimento. Como punto de arranque decidimos construir un único modelo de Markov que llevase a cabo al mismo tiempo las tareas de delimitación y clasificación. Por tanto no se aplica ninguna transformación al corpus ni se utiliza ningún clasificador adicional.

Los resultados no son malos, aunque sí se observa cierta debilidad en la categoría MISC debido seguramente a que es la menos natural de las definidas en la ontología de CoNLL.

	precisión	cobertura	$F_{\beta=1}$
PER	80.04 %	56.19 %	66.03 %
LOC	77.39 %	67.90 %	72.33 %
ORG	73.28 %	69.93 %	71.56 %
MISC	46.37 %	33.82 %	39.12 %
Total	73.52 %	63.02 %	67.87 %

Tabla 1: Delimitación y clasificación con modelos de Markov

La tabla 2 muestra los resultados de la tarea de delimitación (sin clasificación) de entidades . En el experimento *TrEt* se transforma el corpus de entrenamiento previamente a la estimación del modelo de Markov, cambiando las etiquetas tal y como se indica en la figura 2. En el experimento *TrEtVc* se transforman tanto el corpus de aprendizaje como el de test, cambiando las etiquetas y el vocabulario tal y como se muestra en la figura 3.

Los resultados muestran que la utilización de los modelos de Markov exclusivamente en la delimitación conlleva una sensible mejora con respecto al experimento inicial (tabla 1). Esta mejora es bastante más acusada en el experimento *TrEtVc* donde la estimación del modelo se ve beneficiada por el aprovechamiento de la información que proporcionan las mayúsculas, las palabras delimitadores y la reducción del tamaño del vocabulario.

	precisión	cobertura	$F_{\beta=1}$
TrEt	79.98 %	66.56 %	72.66 %
TrEtVc	87.21 %	84.71 %	85.95 %

Tabla 2: Sólo delimitación con modelos de Markov

4.2. Experimento 2

En este segundo experimento partimos del mejor resultado en el proceso de delimitación (experimento 1, *TrEtVc*). Para cada una de las entidades detectadas se genera un vector de características. La base de datos resultantes sirve como conjunto de entrenamiento para la obtención de un árbol de decisión. La tabla 3 muestra los resultados de cinco pruebas distintas, que difieren en las características utilizadas para generar cada vector:

- Base: se utilizan características ortográficas, contextos, sufijos y palabras significativas. Estas características se utilizan como base en el resto de modelos.
- RedCont: se reducen las listas de contextos y palabras significativas en un 25 %.
- RefMisc: se refuerza la información sobre la categoría MISC. Para ello se generan nuevas listas de palabras significativas que incluyen aquellas palabras relevantes para MISC y no relevantes para el resto de categorías.
- Ext: se incorporan características calculadas en base a listas externas (ciudades, nombres de persona, apellidos, ...).
- Todas: se incluyen al experimento Base, las características de los otros tres experimentos.

Las tres modificaciones propuestas mejoran por separado los resultados del experimento base. La mejora más significativa se debe a la reducción de un 25 % en las listas

	precisión	cobertura	$F_{\beta=1}$
Base	63.44 %	61.62 %	62.51 %
RedCont	67.75 %	65.81 %	66.76 %
RefMisc	66.68 %	64.77 %	65.71 %
Ext	64.71 %	62.85 %	63.77 %
Todas	68.93 %	66.96 %	67.93 %

Tabla 3: Clasificación con árboles de decisión

de palabras significativas y contexto, con ello se consigue relajar la dependencia de los ejemplos del corpus con respecto a estas características lo que impide que afecten de forma negativa a los ejemplos del test que no las presenten. En menor medida mejoran el experimento base el refuerzo de la categoría MISC, que orienta el sistema a la categoría *más difícil*, y la utilización de información externa al corpus de aprendizaje.

Con la combinación de las tres modificaciones se alcanza un resultado similar al obtenido en el experimento inicial (tabla 1) en el que se utilizaba un modelo de Markov para reconocer y clasificar al mismo tiempo. No obstante, queda la impresión de que aún hay margen de mejora ya que la suma de las mejoras obtenidas por separado con cada una de las modificaciones superaría sensiblemente el resultado de la combinación. Esta impresión quedará confirmada en el siguiente experimento, en el que utilizando técnicas de aprendizaje más complejas se consigue sacar más partido de los datos utilizados en éste.

4.3. Experimento 3

En este experimento se comprueba el efecto que tiene la utilización de técnicas que generan distintos clasificadores a partir del mismo conjunto de entrenamientos. Se comparan los resultados obtenidos con el mejor modelo del experimento 2 en el que se utilizan árboles de decisión como modelo de aprendizaje con los obtenidos aplicando *boosting* y *bagging*.

	precisión	cobertura	$F_{\beta=1}$
Árbol	68.93 %	66.96 %	67.93 %
Boosting	70.12 %	68.11 %	69.10 %
Bagging	71.10 %	69.06 %	70.07 %

Tabla 4: Clasificación con *boosting* y *bagging*

Se han obtenido mejoras en los resultados en ambos casos, lo que muestra la importancia de la elección de un modelo de aprendizaje apropiado en la etapa de clasificación. Esta es una de las líneas que queremos explotar en el futuro, experimentando con clasificadores más versátiles como las SVM (*Support Vector Machines*) o las propias técnicas de *bagging* y *boosting* aplicadas a clasificadores más simples, que se han demostrado muy útiles en este tipo de problemas (Carreras et al., 2002).

5. Conclusiones y trabajo futuro

El sistema presentado en este trabajo constituye una primera aproximación al problema de Reconocimiento de Entidades con Nombre. Los resultados obtenidos situarían a nuestro sistema en la parte media del *ranking* de los trabajos presentados en la tarea NER de CoNLL 2002.

La poca relevancia de la información externa al corpus hace que el sistema sea fácilmente portable ante cambios de idioma y de conjuntos de categorías. La división entre delimitación y clasificación facilita el desarrollo y la mejora del sistema, ya que permite atacar ambos problemas de forma separada usando distintas técnicas. Para el problema de la delimitación, los modelos de Markov se han mostrado como una técnica bastante útil, ya que con unas simples transformaciones del corpus se han obtenido unos resultados muy buenos. En la etapa de clasificación el rango de técnicas a utilizar es muy amplio, hemos comenzado utilizado árboles de decisión con unos resultados aceptables, pero una vez obtenida una base de datos a partir del corpus podemos experimentar con cualquier tipo de clasificador.

En nuestra agenda de trabajos futuros hay muchos frentes abiertos por los que continuar nuestra investigación. Con respecto a la delimitación aún se puede afinar más experimentando con nuevas características que permitan agrupar las palabras del vocabulario mediante un criterio lingüístico en lugar de con simples expresiones regulares. En la parte de clasificación, las vías de trabajo pasan por la búsqueda de nuevas características que aporten más información a los clasificadores y la experimentación con nuevos modelos de aprendizaje. El experimento 3 ha sido un primer intento en este sentido, y nos ha abierto un camino esperanzador mostrándonos que se puede sacar más partido de los mismos datos de entrenamiento eligiendo la técnica de aprendizaje apropia-

da.

Bibliografía

Brants, T. 2000. TnT - a statistical part-of-speech tagger, En *Proceedings of the 6th Applied NLP Conference (ANLP00)*, páginas 224–231, Seattle, USA.

Carreras, X., L. Màrquez y L. Padró. 2002. Named Entity Extraction using AdaBoost, En *CoNLL02 Computational Natural Language Learning. Shared Task*,

Chinchor, N. 1998. MUC-7 Named Entity Task Definition (Version 3.5), En *Message Understanding Conference Proceedings MUC-7*, Fairfax, Virginia, USA.

Mannig C. y H. Schütze. 1999. *Foundations of Statistical Natural Language Processing* MIT Press.

Rössler M. 2002. Using Markov Models for Named Entity recognition in German newspapers, En *Proceedings of the Worl¡kshop on Machine Learning Approaches in Computational Linguistics*, páginas 29–37, Trento, Italia.

Tjong K. S. E. 2002. Introduction to CoNLL-2002 shared task: Language-independent named entity recognition En *Proceedings of Sixth Conference on Natural Language Learning CoNLL 2002*, Taipei, Taiwan.

Witten I. H. y E. Frank. 2000. *Data Mining*, Morgan Kaufmann Publishers.

Zhou, G. y J. Su. 2002. Named Entity Recognition using an HMM-based Chunk Tagger, En *Proceedings of ACL 2002*, páginas 473-480